


Data-driven federated learning in drug discovery with knowledge distillation

Received: 29 February 2024

Accepted: 14 January 2025

Published online: 05 March 2025

 Check for updates

Thierry Hanser¹✉, Ernst Ahlberg^{2,3}, Alexander Amberg⁴, Lennart T. Anger^{4,5}, Chris Barber¹, Richard J. Brennan⁶, Alessandro Brigo⁷, Annie Delaunois⁸, Susanne Glowienke⁹, Nigel Greene^{10,11}, Laura Johnston¹, Daniel Kuhn¹², Lara Kuhnke¹³, Jean-François Marchaland¹, Wolfgang Muster⁷, Jeffrey Plante¹, Friedrich Rippmann¹², Yogesh Sabnis⁸, Friedemann Schmidt⁴, Ruud van Deursen^{1,14}, Stéphane Werner¹, Angela White¹⁵, Joerg Wichard^{13,16} & Tomoya Yukawa¹⁷

A main challenge for artificial intelligence in scientific research is ensuring access to sufficient, high-quality data for the development of impactful models. Despite the abundance of public data, the most valuable knowledge often remains embedded within confidential corporate data silos. Although industries are increasingly open to sharing non-competitive insights, such collaboration is often constrained by the confidentiality of the underlying data. Federated learning makes it possible to share knowledge without compromising data privacy, but it has notable limitations. Here, we introduce FLuID (federated learning using information distillation), a data-centric application of federated distillation tailored to drug discovery aiming to preserve data privacy. We validate FLuID in two experiments, first involving public data simulating a virtual consortium and second in a real-world research collaboration between eight pharmaceutical companies. Although the alignment of the models with the partner specific domain remains challenging, the data-driven nature of FLuID offers several avenues to mitigate domain shift. FLuID fosters knowledge sharing among pharmaceutical organizations, paving the way for a new generation of models with enhanced performance and an expanded applicability domain in biological activity predictions.

Artificial intelligence's (AI's) main power in solving scientific problems lies in its capability to extract knowledge from data by mining causal patterns from experimental observations. This form of digital knowledge is used to build powerful predictive models and has become a central component of the modern scientific toolkit. For instance, by analysing a large number of medical lung images, AI algorithms have identified causal relationships between photographic features and the probability of a patient being diagnosed with cancer¹. AI can outperform human experts in detecting concerning anomalies and allows early treatment of patients². With the assistance of AI, scientists embrace

new powerful tools and strategies to solve problems by harnessing the valuable knowledge embedded in data.

Often, the learning process requires large amounts of data. The availability of sufficient high-quality data depends largely on the domain and ultimately determines the real impact of AI. Therefore, one of the main challenges in AI is to access such pivotal data collections; a task particularly difficult in domains where intellectual property and data governance are paramount. This limitation prevents organizations from sharing and accessing a wealth of knowledge that remains locked in private silos. The private nature of data, due to either

A full list of affiliations appears at the end of the paper. ✉e-mail: thierry.hanser@lhasalimited.org

corporate or personal confidentiality requirements, has become a major bottleneck in leveraging new AI tools and limits their benefits. Unlocking the knowledge embedded in private data would dramatically augment the impact of AI and open an avenue to a new generation of predictive models with improved performance and wider applicability domain (AD). This perspective has led to intensive research in the field of federated learning (FL) currently dominated by model-centric approaches. However, despite the enthusiasm surrounding FL, it is important to acknowledge their limitations from a privacy perspective and the ethical considerations they entail. Recent discussions highlight important concerns regarding data privacy and governance in FL frameworks^{3,4}, particularly in the context of digital health⁵. The inherent heterogeneous and biased nature (not independent and identically distributed⁶) of data across organizations sharing knowledge is an additional challenge for FL, especially when a single shared model is used to fit all the local data conditions. These challenges underscore the necessity of moving beyond simple model-driven FL approaches to effectively address privacy and ethical considerations and account for the frequent non-homogeneous distribution of data.

MD-FL

FL was introduced by Google in 2016 (ref. 7) and has since drawn a lot of attention in the world of AI^{8–13}. This research effort aims to build better models and is driven by multiple facets including the benefit of decentralizing the learning process across an ensemble of devices (cross-device FL), accessing widely distributed knowledge (cross-silo FL) and preserving privacy of local data. At present, the most adopted FL approach is model driven¹⁴. It is based on a central model, trained across delocalized sites. The knowledge is extracted at each site through local model training and parameter updates are federated into the central model via secure network communication. The central model becomes the recipient of the federated knowledge, and the parameter updates convey the knowledge from the local site to the central model (Fig. 1a). Since only the model parameters are shared, the method provides a good level of privacy. This model also distributes the training effort across different local computing resources and therefore scales well. Several techniques allow further preservation of privacy, for instance by introducing carefully calibrated noise at different stages of the process, using differential privacy^{15,16}. Model-driven FL (MD-FL) is a powerful approach that is already applied across a wide range of domains, including finance¹⁷, healthcare^{18–20} and agrochemical and drug discovery^{21–26}. Despite its benefits and successful adoption, this approach also has several limitations and its implementation can be challenging. First, exposing a shared model directly trained on private and potentially personal data to third-party organizations may leak some level of privacy and therefore be exposed to all the challenge related to data governance. Second, MD-FL requires a non-trivial protocol and infrastructure to orchestrate the training of the local and central models and ensure secure communication between the corresponding sites. Next, the network traffic between the local and the central models can become substantial. The learning process often involves several rounds of model updates, and these cycles require further orchestration and can span a long period of time. Furthermore, the central model architecture must be predefined and frozen, changing the model's configuration would invalidate previously learned parameters preventing future optimization. Finally, the extracted knowledge is trapped in a static model designed for a specific use case and offers very limited repurposing options.

The communication overhead in MD-FL can be partially mitigated using weight update compression techniques²⁷ and more recently knowledge distillation²⁸ (KD) has been used to further improve model-centric FL. MD-FL benefits from KD by reducing the communication overhead and enhancing the privacy of models. This latter improvement is mainly achieved by locally and privately distilling knowledge from a complex teacher model to a smaller shared student,

which requires fewer weight updates and hence significantly reducing communication overheads. The distillation process also prevents direct access to the teacher model trained directly with private data and only exposes the student model as a shared resource. Despite this positive synergy between KD and MD-FL, the model-centric paradigm still suffers from limitations due to the rigidity of the shared model architecture and the static final model format that cannot be easily repurposed.

DD-FL

To overcome the limitations of MD-FL in the context of drug discovery, we propose to use data-driven-FL (DD-FL). Instead of sharing parameters of a central model, DD-FL relies on sharing and consolidating annotations of non-sensitive surrogate data (Fig. 1b). Whereas in MD-FL the knowledge recipient is a central model fed by parameter updates, in DD-FL, the recipient is a large public transfer dataset and the knowledge is transferred by predicted labels. DD-FL leverages the benefits of semisupervised learning²⁹ (label propagation), ensemble prediction (label consolidation) and KD (soft labels). The resulting federated knowledge is captured in a universal and perennial format (annotated data); it can be easily used in many different machine learning (ML) contexts and holds even more benefits through future innovation in these domains. Furthermore, since the knowledge is stored in the form of labels, it is decoupled from the underlying federation method that can be continuously improved without invalidating previous learning efforts. Another important trait of DD-FL is the possibility of using different learning algorithms with bespoke local configuration for each participant. Indeed, there is no coupling between a local and a central model since the knowledge is conveyed by labels rather than model parameters; DD-FL supports heterogeneous FL. Finally, DD-FL is by nature incremental, new federated data originating from future learning rounds can be directly merged to the existing asset; it becomes possible to dynamically expand the knowledge domain and support active FL. Active FL can be used to augment the density of the transfer data in regions of the application space requiring more knowledge. In this article we describe federated learning using information distillation (FLuID) a mature implementation of DD-FL in the context of drug discovery and toxicity prediction. Leveraging DD-FL, the approach not only facilitates the extraction and transfer of knowledge while preserving data privacy but also enables FL without direct access to sensitive data. By using surrogate data for general knowledge transfer, FLuID protects confidential corporate information in an industrial context and helps mitigating ethical concerns associated with personal data usage for instance in the context of digital health.

The method uses knowledge transfer and KD³⁰ and is based on the teacher–student approach^{31,32} adapted to facilitate knowledge sharing in the domain of drug discovery. The teacher–student approach is a two-step knowledge transfer method. First a private teacher model is trained from the private data. The teacher model is not exposed directly to end users, instead, it is kept private and used to annotate public data. The annotated yet non-sensitive dataset, called the transfer data, can subsequently be used to train a secondary model called the student model that can be made accessible to the end user. The method's underlying indirection protects the original proprietary data from being directly exposed to the end user and hence preserves its privacy. Further privacy can be achieved by using an ensemble of teachers³² where the transfer data are split into several subsets obfuscating even more the original dataset. Our work is based on the idea that confidential corporate data form a natural collection of subsets from which it is possible to build an ensemble of private teachers and apply KD. Hence a second important aspect of the proposed method is the ability to federate the knowledge extracted from several sources of private data. This federation is achieved by using the same public transfer data to extract knowledge across the ensemble of private teachers. For each public data instance, the set of labels predicted by

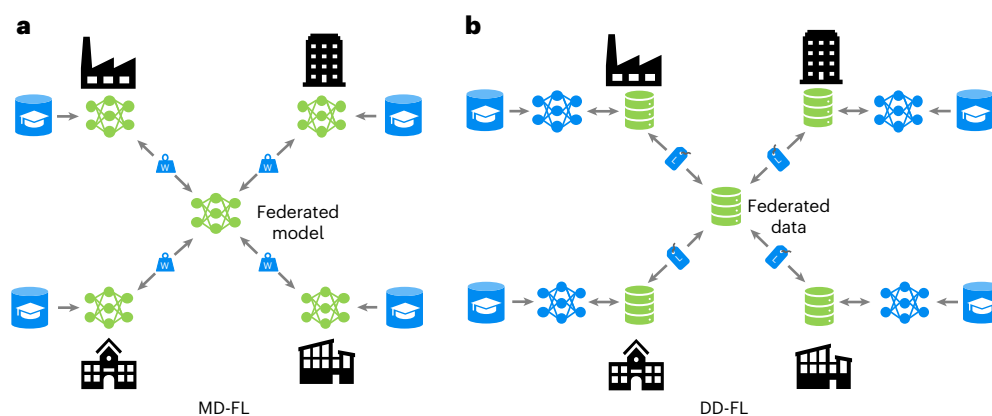


Fig. 1 | FL schemes. a, MD-FL where a central model is trained in a delocalized manner using local private data; knowledge is conveyed through local weight updates. The weight updates are aggregated in the central model. **b**, DD-FL

relying on a central public dataset; knowledge is conveyed via labels predicted by local private models. The ensemble of labels for each public instance are consolidated in the central dataset.

the private teachers is consolidated into a single label. The resulting federated labels benefit from the contributions of an ensemble of teacher models^{33,34} and take advantage of the knowledge domains across all the corresponding data sources. The hypothesis driving the FLuID approach is that a student benefiting from the education of several teachers will become more knowledgeable than any individual teacher. By extracting and federating the knowledge embedded in many separate private data silos, it is possible to create models that are more performant than models built from a single source of data. In other words, it is possible for several organizations to collaborate and share knowledge to create better models without disclosing any private information.

Related work

Recently, knowledge distillation has been combined with FL, leading to a new research direction ‘federated distillation’ (FD)^{35,36}. FD was initially motivated by the need of reducing the communication overhead induced by model-centric approaches and has quickly been adopted to also improve the privacy of knowledge transfer and mitigate the heterogeneity of the different sources of data allowing local model configurations. The concept of FD was successfully applied in the context of image classification^{37–39} and medical relation text extraction⁴⁰ for instance. Several variations of FD have emerged integrating different other deep learning techniques such as contrastive learning⁴¹, and generative adversarial networks to actively augment the transfer data when needed⁴². FD is not the only possible architecture for DD-FL and several methods involving different architectures^{43,44} have been developed recently.

With the current work, we demonstrate in a real industrial setup the potential of data-centric FL and its application in the context of drug discovery, namely, the improvement of quantitative structure activity (QSAR) models.

FLuID

FLuID introduces an implementation of FD in the context of drug discovery. We describe in detail the corresponding implementation of DD-FL and present a proof of concept using the prediction of biological properties for chemical compounds as an exemplifying task. For transparency purpose, the method evaluation is based on public data and available as open-source software⁴⁵. Furthermore, we show how the approach was successfully applied in a real-world context to extract, transfer and federate hERG (human ether-a-go-go gene-encoded ion channel)⁴⁶ activity knowledge from tens of thousands of drug discovery research compounds across several large pharmaceutical companies. We also show that FLuID can improve the performance of hERG activity

models and augment their AD. It is important to note that this use case does not involve personal data and is therefore not subject to data governance restriction such as the European General Data Protection Regulation⁴⁷. However, such concerns need to be carefully addressed in contexts involving personal data such as medical data. In such cases, synthetic surrogate data can be generated using fictive individual profiles. Using fully anonymized surrogate data and KD facilitates this compliance. The presented work will assume that the surrogate data are fully anonymized and will focus on the method’s ability to improve the models’ performance and AD.

FLuID is domain agnostic. We describe and explain its principle in the context of biological activity prediction where the data are chemical compounds associated with experimental activity labels. We chose the important hERG binding classification as the predictive task to describe and validate the methodology, as the interaction of a drug candidate with the hERG channel can lead to arrhythmias and cause sudden death.

The method is divided in three main phases (Fig. 2), which are discussed in detail in the Methods section:

- (1) Knowledge extraction: knowledge is extracted from private data using ML algorithms to train a model called the teacher. This model is used to annotate a large set of public chemical compounds, transferring knowledge without revealing confidential information. The annotated transfer data must be diverse, tractable and homogeneous to effectively represent a useful chemical space. A well-designed transfer dataset ensures broad coverage and successful knowledge transfer, enabling organizations to share knowledge without compromising privacy.
- (2) Knowledge consolidation: knowledge from multiple partners is consolidated into a federated dataset. In this process, each teacher has previously predicted a label for each public compound. This ensemble of predictions per compound is consolidated into a single label using a selected label merging policy. The resulting federated labels captures the knowledge from all contributing organizations and can be used to build a federated model called the student model.
- (3) Knowledge fusion: the federated student model, built from the consolidated dataset, can be directly integrated into partners’ internal processes. FLuID also allows partners to combine federated data with their private data to build hybrid models. This flexibility enables partners to select the most relevant data and ML algorithms for specific tasks, adjusting for domain needs or performance goals.

By leveraging FD, this methodology enables the transfer of valuable insights from private datasets and consolidates knowledge across

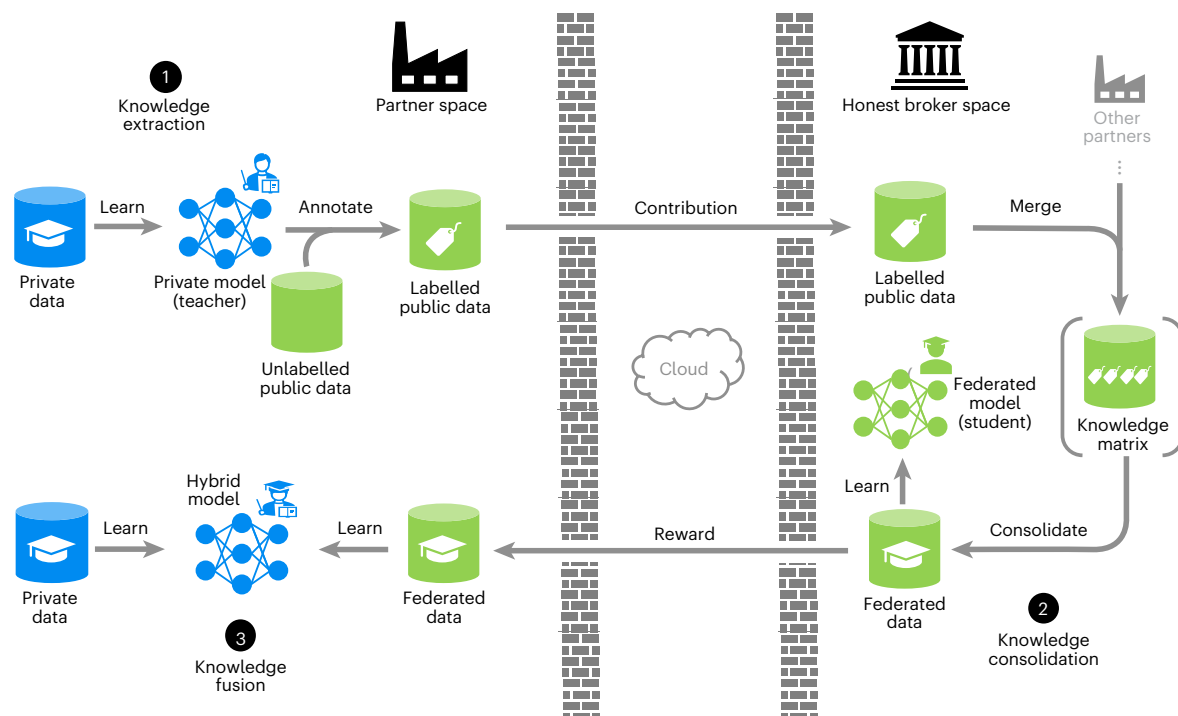


Fig. 2 | FLuID methodology. (1) First the knowledge is learned from the private data in the form of a private model trained with this proprietary data; the resulting model is called the teacher. The knowledge captured by the teacher model is then transferred and stored in non-sensitive data by annotating a large collection of public structures with the corresponding activity predicted by the teacher. (2) Once all participants have completed the knowledge extraction phase and contributed a label for each public compound, this ensemble of labels is consolidated into a single more robust federated label. Next, the

resulting collection of public structures with consolidated labels is used to train a secondary model called the federated student. This model indirectly learns and uses the knowledge contained in the private data of all the participating organizations without directly exposing it and can be shared with no risk of disclosing confidential information. (3) To maximize model improvement, partners can now collect and fuse the federated data with their own available private (and potentially public) data to build a hybrid model.

multiple organizations while protecting sensitive data. The flexibility to tailor the federated dataset and the ability to integrate the knowledge into existing workflows allows for optimization across various use cases and domains. This approach fosters collaboration and enables organizations to build powerful hybrid models and enhance predictive capabilities while maintaining data privacy.

Privacy considerations

Privacy preservation is paramount to alleviate confidentiality concerns and to encourage knowledge sharing through FL. FLuID does not expose either the private data or the teacher model outside the private space of the contributing organizations. Instead, only the predicted annotations for public structures leave this space to be federated. Although each prediction has in theory a minimum privacy cost⁴⁸ the extremely reduced information contained in the annotation of a public compound in regard to the original private data indicates that a malicious attempt to rebuild information about the original private space would require an extremely large number of specifically designed queries against the student model to get a limited understanding of the nature of this space. The fact that the number of labels is restricted (in our example to 350,000) and that these compounds have been selected randomly and a priori, ensures that the privacy cost is constant and close to zero. Intuitively, one can conclude that it is impossible to rebuild the teacher's training data by simply exploiting the information contained in the labels predicted for a limited number of public compounds randomly picked in advance. If required, additional obfuscation techniques such as differential privacy¹⁶ can be applied to further reduce and control privacy leak. However, the teacher–student indirection and the very low level of confidential information contained in an annotation mitigates the need of such an extra step. Furthermore, as we have seen, the

teacher labels are merged into a single federated label from which it is impossible to de-convolute back the contribution of each original label. This extra obfuscation further eliminates the risk of tracing back the original private information. By combining surrogate data, KD and label aggregation, the federated dataset moves outside the scope of the data protection, since the data are anonymous and labels cannot be traced back to the original training data. This advantage is a general attribute of FD that shares only unlabelled surrogate data, as is the case for FLuID. We can, therefore, consider the proposed method as a privacy-preserving method to extract, transfer and federate knowledge from private data.

Challenges

We have seen that FLuID addresses many of the FL concerns by introducing a data-driven approach that mitigates ethical and confidential concerns by using surrogate data. The approach is also extremely light in communication and security requirements, as the teacher models are fully trained locally and only non-sensitive data annotations (labels) are shared in a single round. Despite this promising outlook, there are remaining challenges:

- The annotations of the surrogate data by the teacher are subject to epistemic noise introduced by the teacher model. Assuming this model is not perfect, the labels contributed by a partner contains noise from both the original data and the teacher trained with this data. This noise can be important if the partner does not possess good quality data. However, this noise can be partially mitigated at consolidation time by the beneficial effect of having an ensemble of teachers predicting each surrogate data point; the consolidated label is therefore more robust. Additionally, the soft

- labels produced by the teachers and consolidated into a single soft label are more informative and can be exploited when using the federated data to train future models to further mitigate noise⁴⁹.
- In FD, label bias in the initial private datasets can result in teachers that perform poorly, which introduces noise into the aggregated federated data despite mitigation efforts. One approach to counteract this is to conduct multiple federation rounds, allowing participants to balance their private data with federated data shared in previous rounds. This iterative adjustment helps to attenuate the label bias by progressively integrating additional diverse, representative samples into each partner's dataset. Similarly, feature-level bias in the private data can be partially addressed during surrogate data preparation. By enforcing a fair distribution of instance attributes (for example, sampling from an even or stratified distribution), surrogate data can serve as a buffer against the underlying bias in private datasets. This method reduces the extent to which skewed distributions in the original data influence the final model, enabling surrogate data to act as an 'anti-bias' mechanism and preserving model performance across heterogeneous sources.
 - One of the most challenging problems in ML is domain shift, that is, the gap between the domain of the data used to train the model and the domain where the model is applied⁴⁵. In FL, partners often operate in different domains that can amplify the negative effect of domain shift. The issue is not specific to FL or FLuID and affects ML in general. However, the variety of domains contributed by the partners may favour a wider AD of the resulting models that can be an advantage when a partner needs to explore new domains not yet covered by their own private data. Furthermore, in FLuID, the federated data can be tuned to target a specific domain before its fusion with the private data. When training the hybrid model, FLuID offers more flexibility than model-driven approaches to apply further domain adaptation techniques to mitigate domain-shift issues.
 - As for any ML approach, it remains important to use homogenous data with respect to the target task. This can be especially challenging in FL where different partners may operate in different ways. Homogeneity can be achieved by aligning the data compilation, curation and validation protocols across all the partners. Often the protocol agreement and implementation steps can be lengthy and resource intensive. In the case of this work, we have focused on single prediction tasks (hERG activity), carefully selected compatible data sources (biological assays) and used the same classification thresholds across all partners. This alignment exercise has led to very interesting and useful discussions across the partners' scientific community and was perceived as a positive byproduct of this research.

Validation

To validate the FLuID method, we must demonstrate that the knowledge contained in the private data can be successfully extracted by the teacher model, stored in the transfer data and subsequently learned by the student model. To quantify this knowledge transfer, we compare the performance of the student model with the performance of its individual teachers using an external reference test set. Our validation criterium is expressed as follows: 'if the student is performing at least as well as the average teacher, then knowledge was successfully extracted and transferred from the private data to the student model' (equation (3)).

Given the average teacher performance $\overline{\text{Score}}_t$:

$$\overline{\text{Score}}_t = \frac{1}{P} \sum_{i=1}^P \text{Score}_{t_i} \quad (1)$$

and the student performance Score_s , the validation criterium can be expressed as:

$$\text{Score}_s \geq \overline{\text{Score}}_t \quad (2)$$

where P is the number of partners in the federation consortium. We used the Mathew correlation coefficient⁵⁰ (MCC) value as our primary metric (score) to objectively compare the performances of models since this metric is global (covers precision and recall aspects) and remains robust even in the case of imbalanced test sets. To fully validate the FLuID method, we performed two similar validation experiments. In the first experiment, we used public data to simulate a virtual consortium and implemented the method as a transparent and reproducible open-source Python notebook⁴⁵. For the second experiment, we initiated a research collaboration across eight pharmaceutical companies to validate the methodology in a real industrial context. Details of the validations are described in the Methods section.

Public proof-of-concept results

The results for the consortium simulation using public data are summarized in Fig. 3a and 6. As expected, the performance of a teacher directly depends on the size, bias and chemical space overlap of its corresponding cluster data compared to the test set, D_{test} . This explains the disparity of the teacher performance with an MCC value ranging from 0.151 to 0.486. The average teacher performance, $\overline{\text{MCC}}_t$, is 0.320. The federated student trained with only the federated data exhibits an MCC of 0.551, which means that our success criterium has been met ($0.551 > 0.320$).

Not only does the federated student performance exceed the average teacher performance, but the student outperforms all the individual teachers that it learned from (Fig. 3b). We can conclude that valuable knowledge contained in the private data has been successfully transferred to the student, without exposing any confidential information as the student has only been exposed to the federated data. Furthermore, in this experiment, the federation process allowed the student to become more knowledgeable than any individual teacher, which indicates a positive synergistic effect of this process. It is noteworthy that the federated student model has never directly seen any experimental data and yet outperforms the models built using the experimental data. This achievement further demonstrates the actual value of the knowledge extraction and federation using the data-driven FLuID approach. We also observe that the student model outperforms the average teacher in all calculated performance metrics, not just MCC, indicating that the proposed method leads to useful and robust federated knowledge.

Contribution of each teacher

We measured the relative contribution of each partner to the federated knowledge. We wish to ensure that all teachers are equitably participating in the consolidation process. To disentangle and measure the contribution of each individual teacher we computed the average contribution weight for each teacher during the label consolidation step. We observed a fair contribution of each teacher to the final labels (Fig. 3c) indicating that each source of private data has contributed knowledge to the final federated data.

Leave one teacher out

Even though all teachers contribute to the federated labels, the quantitative contribution does not capture the quality of the knowledge or if, instead, noise was introduced in the federation process. To measure the contribution of each teacher with respect to the performance of the student, we have repeated the validation experiment eight times and left a different teacher out for each iteration (Fig. 3d).

We did not observe any substantive change in the student performance across the different iterations. The federated student model still performs in a range between $\text{MCC} = 0.449$ and $\text{MCC} = 0.551$ with only seven partners contributing. Therefore, we can conclude that no

a Overall external validation results

Model	Size	BAC	ACC	MCC	F1	SENS	SPEC	PPV	NPV
T1	254	0.603	0.449	0.205	0.449	0.907	0.298	0.298	0.907
T2	258	0.529	0.760	0.151	0.129	0.072	0.986	0.629	0.764
T3	971	0.760	0.790	0.486	0.623	0.701	0.820	0.561	0.893
T4	186	0.566	0.771	0.237	0.259	0.162	0.971	0.645	0.779
T5	1,030	0.628	0.801	0.381	0.414	0.284	0.971	0.764	0.805
T6	1,867	0.625	0.803	0.389	0.407	0.273	0.978	0.800	0.804
T7	281	0.622	0.763	0.288	0.418	0.345	0.900	0.531	0.807
T8	1,371	0.712	0.785	0.424	0.567	0.567	0.857	0.566	0.858
Tmean	777	0.631	0.740	0.320	0.408	0.414	0.848	0.599	0.827
Sfed	10,000	0.766	0.838	0.551	0.655	0.622	0.909	0.693	0.880
H1	10,000	0.766	0.840	0.554	0.656	0.619	0.912	0.698	0.879
H2	10,000	0.761	0.836	0.545	0.649	0.612	0.910	0.691	0.877
H3	10,000	0.777	0.843	0.569	0.671	0.646	0.908	0.698	0.886
H4	10,000	0.770	0.841	0.560	0.662	0.630	0.911	0.698	0.882
H5	10,000	0.765	0.838	0.551	0.654	0.619	0.910	0.694	0.879
H6	10,000	0.771	0.841	0.560	0.663	0.633	0.909	0.696	0.883
H7	10,000	0.764	0.839	0.552	0.654	0.616	0.912	0.697	0.879
H8	10,000	0.786	0.850	0.588	0.685	0.658	0.914	0.714	0.890

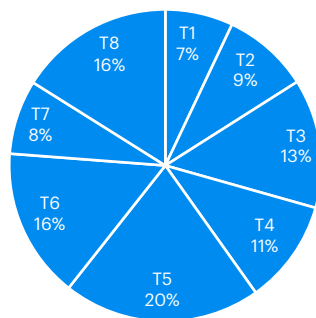
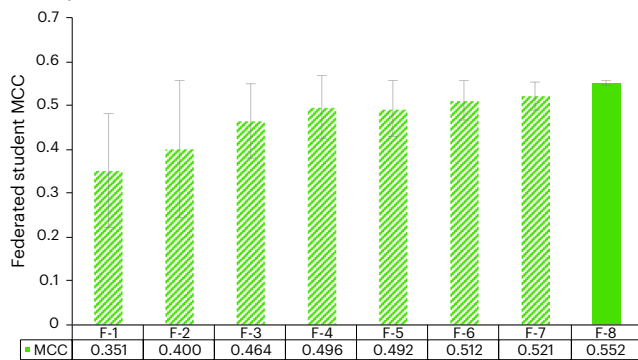
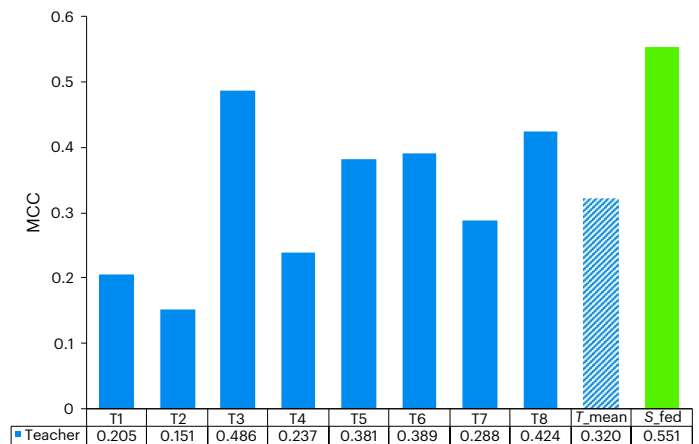
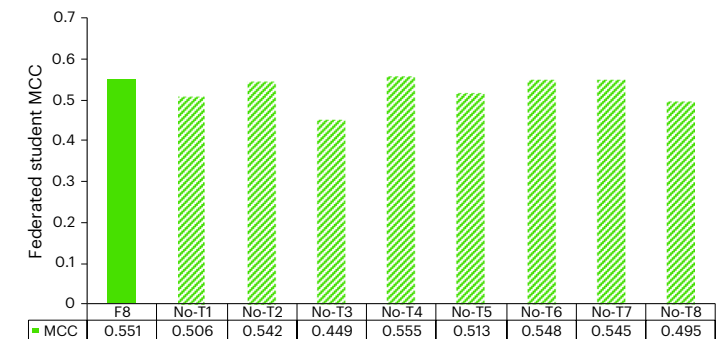
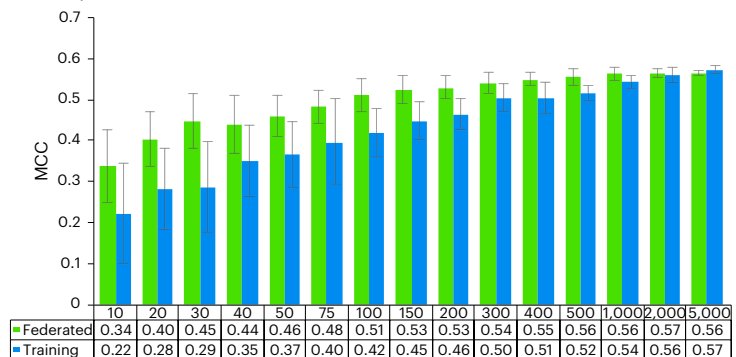
c Teacher contributions**e** Impact of the number of teachers

Fig. 3 | Concept validation and teacher–student comparison for a simulated consortium. a, Performance summary table. **b**, Individual teacher model MCC performance. **c**, Teacher model contributions. **d**, Leave single teacher model out MCC performance. **e**, Contributing teachers' MCC performance (mean

single teacher was an exclusive source of knowledge, and all partners benefit from each other.

Impact of the number of teachers

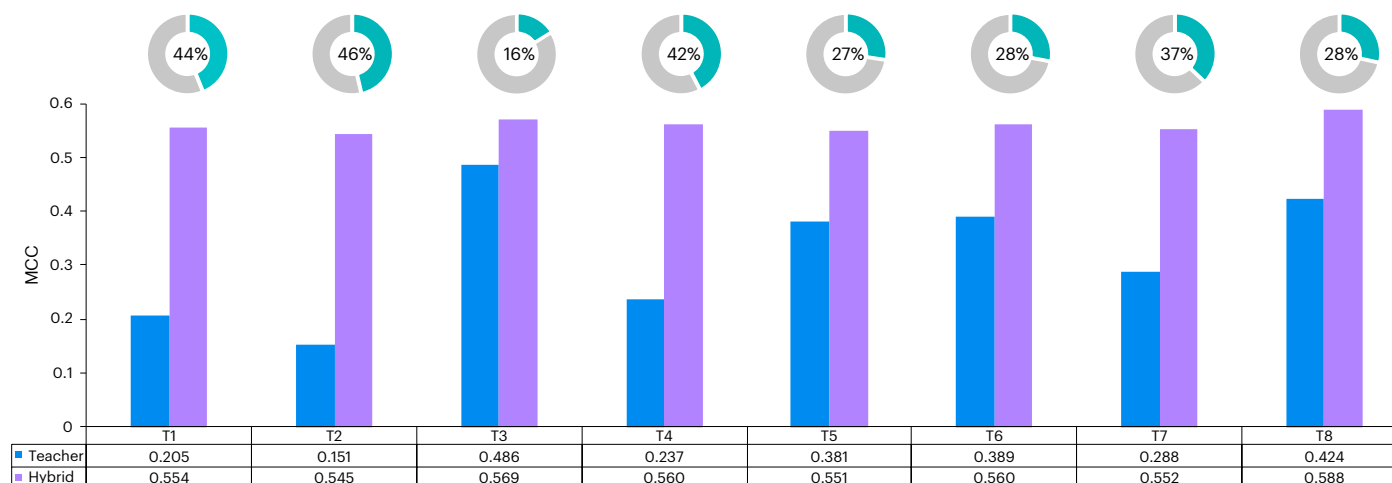
Another important aspect of FL is the size of the consortium. Intuitively, we understand that the benefit of the federation grows with the number of partners; therefore we measured the impact of the number of teachers on the performance of the student model. For this purpose, we repeated the knowledge federation experiment with a variable number of teachers ranging from 1 to 8 (teachers were randomly picked through 30 iterations for each teacher count). As expected, the performance of the student model improves with the number of contributing teachers (Fig. 3e), although it reaches a plateau after four teachers.

b Teacher and student models comparison**d** Leave one teacher out**f** Impact of the student size

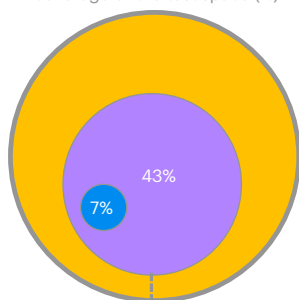
values \pm s.d. $n = 30$). **f**, Student sample size MCC performance (mean \pm s.d. $n = 3$). BAC, balanced accuracy; ACC, accuracy; F1, F1-Score; SENS, sensitivity; SPEC, specificity; PPV, positive predictive value; NPV, negative predictive value.

Size of the student

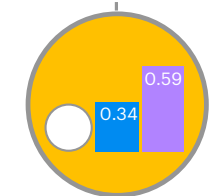
Finally, another interesting aspect of the approach is the amount of the federated data used to build the student model. We measured how the performance of the student evolves with the size of its training set by repeating the model building step with different sizes of random samples from the original federated transfer data (using three replicates for each size). We observed a fast increase of the performance with an early plateau at about 5,000 datapoints (Fig. 3f) indicating that the knowledge is evenly spread and dense across the federated data. Beyond 5,000 datapoints, the performance of the student does not improve significantly. However, the model's AD continues to increase as we present a wider spectrum of structures to its learning algorithm. A slower performance progression with respect to the size of the training

a Model improvement when using federated data

Applicability domain coverage of the test space (%)



New compounds entering AD



Applicability domain performance within the augmented domain (MCC)

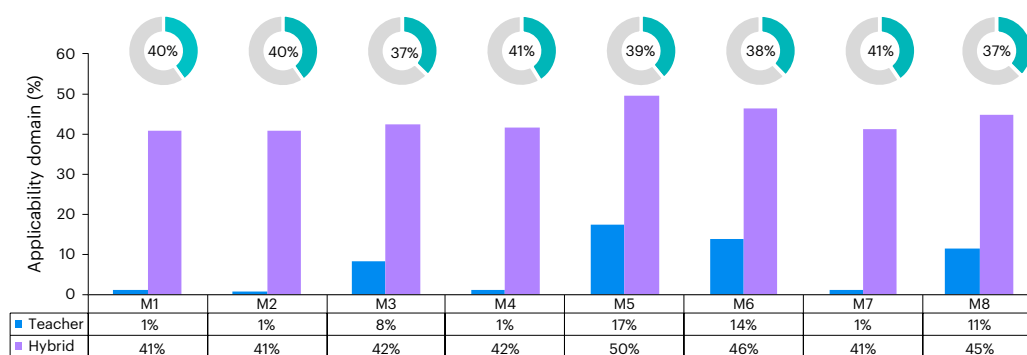
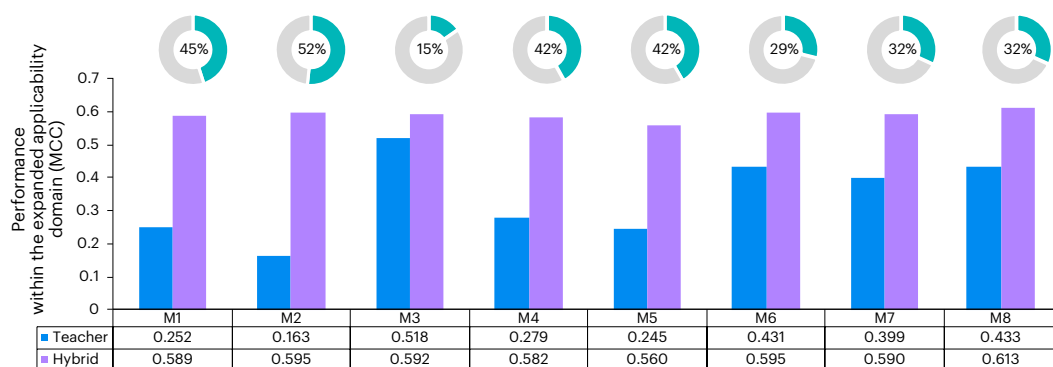
b Applicability domain expansion (on average 39% expansion relative to (a) coverage = 100%)**c** Applicability domain improvement (on average 36% improvement relative to an MCC = 1.0)

Fig. 4 | Teacher and hybrid model comparison. a, Comparison between the hybrid models (private data + federated data) with the teacher models (private data only). **b**, AD expansion through the federated data. **c**, Improvement of performance within the new AD.

set was observed when using the full experimental data. This difference suggests that the distilled federated data are denser in knowledge than the experimental data despite being predicted. This interesting observation opens the possibility that the DD-FL process concentrates the knowledge during the teacher–student distillation. This potential additional benefit of FLuID deserves further investigation.

Improved model predictivity in hybrid models

When we compared the performance of the hybrid models with the corresponding teacher models (Fig. 4a), we again observed an important performance improvement. Each hybrid model significantly outperformed its corresponding teacher model. Combining private and

federated knowledge leads to much better models when compared to models based on private data only. The performance improvement relative to a perfect metric (MCC = 1) varied between 16 and 46%. The average gain in performance was 33%. This substantial improved predictivity across all partners clearly demonstrates the value of DD-FL and its benefit for each individual partner in terms of model performance for the given public benchmark space.

Augmented AD

Finally, we observed an effective augmentation of the AD with increased performance in the newly covered chemical space. The AD coverage significantly increased for each partner (Fig. 4b) and the expansions

a Overall external validation results

	P1	P2	P3	P4	P5	P6	P7	P8	Average teacher	Federated student
MCC	0.539	0.325	0.411	0.270	0.364	0.468	0.321	0.472	0.396	0.546
Kappa	0.529	0.309	0.389	0.256	0.339	0.468	0.255	0.457	0.375	0.499
BACC	0.793	0.634	0.731	0.610	0.643	0.737	0.683	0.704	0.692	0.813
Recall +	0.772	0.353	0.841	0.310	0.350	0.615	0.854	0.473	0.571	0.911
Recall -	0.814	0.916	0.722	0.910	0.937	0.858	0.511	0.934	0.825	0.714
Precis +	0.582	0.584	0.472	0.535	0.650	0.593	0.370	0.707	0.562	0.517
Precis -	0.914	0.808	0.893	0.797	0.811	0.869	0.913	0.841	0.856	0.960
Coverage	0.587	0.392	0.454	0.328	0.200	0.442	0.307	0.706	0.427	0.774

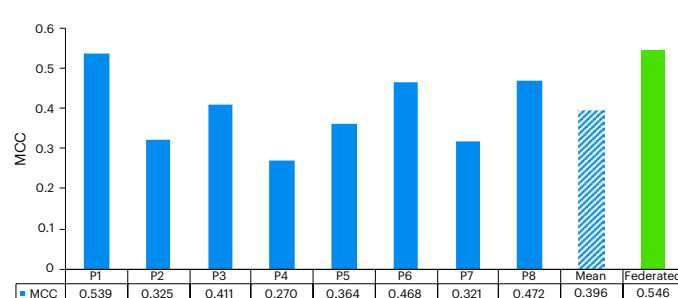
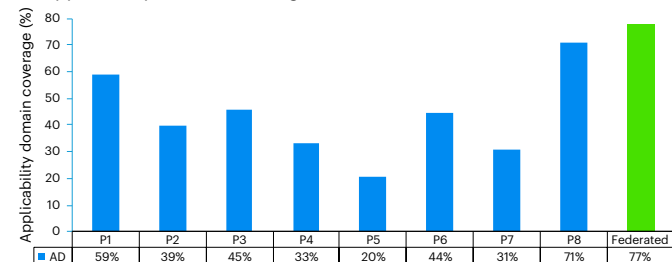
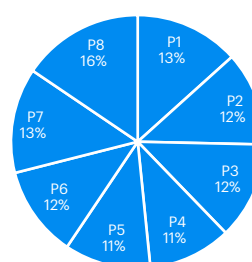
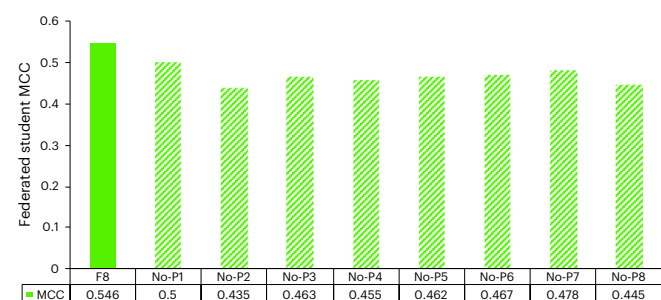
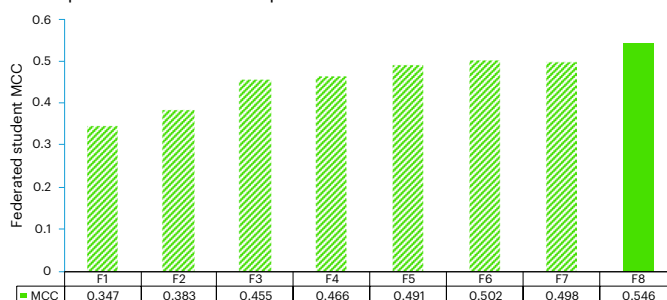
b Teacher and student model comparison**c** Applicability domain coverage**d** Partner contributions**e** Leave one teacher out**f** Impact of the number of partners

Fig. 5 | Concept validation in an industrial setup. a, External partner performance summary table. **b**, Individual partner performance. **c**, Partner AD. **d**, Individual partner contributions. **e**, Leave one partner out performance. **f**, Increasing partner count performance.

relative to a perfect coverage (coverage 100%) range from 37 to 41% with an average of 39% across all partners. When we compare the performance of the teacher models and the hybrid models within the expanded AD subspace, we observe an improvement in performance from 15 to 52% with an average of 36% across partners (Fig. 4c). This simultaneous expansion of the AD and the improvement of the model's predictivity inside this new AD demonstrate that the participants all benefit from sharing knowledge and can access high-performance models with augmented AD.

Industrial results

Alongside validating the methodology with a simulated virtual federation consortium, we also performed a real-world application of the FLuID approach. The industrial validation experiment focused on demonstrating that the knowledge extraction and federation was successful and was therefore centred on the performance of the student model against a publicly available test set. Future experiments will also include the validation of hybrid models and include more endpoints.

When running the proof-of-concept experiment in a real consortium of eight large pharmaceutical partners (P1–P8) we were able to confirm all the trends observed during the simulation exercise. The full teacher and student model performances are presented in Fig. 5a. In this real-world context the federated student model had a predictive performance (MCC = 0.546) greater than the average teacher (MCC = 0.396)

meeting the validation criterium. The student trained only with federated annotations outperformed all the teacher models trained with private experimental data (Fig. 5b). We can conclude that FLuID allows for successful extraction and federation of hERG activity knowledge from the eight industrial partners. The value of this federated knowledge was confirmed through good performance of the corresponding student model. The student model also exhibited a wider AD (AD = 77%) compared to individual teacher models (20–71%) as shown in Fig. 5c. The student models' performances suggest that federated data annotation embeds knowledge comparable to experimental data while offering a very large size (350,000 annotated compounds) covering a wide AD.

When comparing the contribution of each industrial partner, we observed a homogenous participation of each partner indicating that the federated data captured knowledge from all the participants (Fig. 5d). The performance impact of each individual partner was measured in the leave one teacher out validation (Fig. 5e) and no single partner stands out as a critical contributor, indicating a fair reward for all the participants.

Finally, in the industrial setup, we observe again a fast increase in the performance of the federated student models as the consortiums growing from one to four partners (Figs. 5f). After five or more partners, we note a plateau in predictive performance, but the AD grows as more sources of knowledge from different chemical spaces feed the federation process.

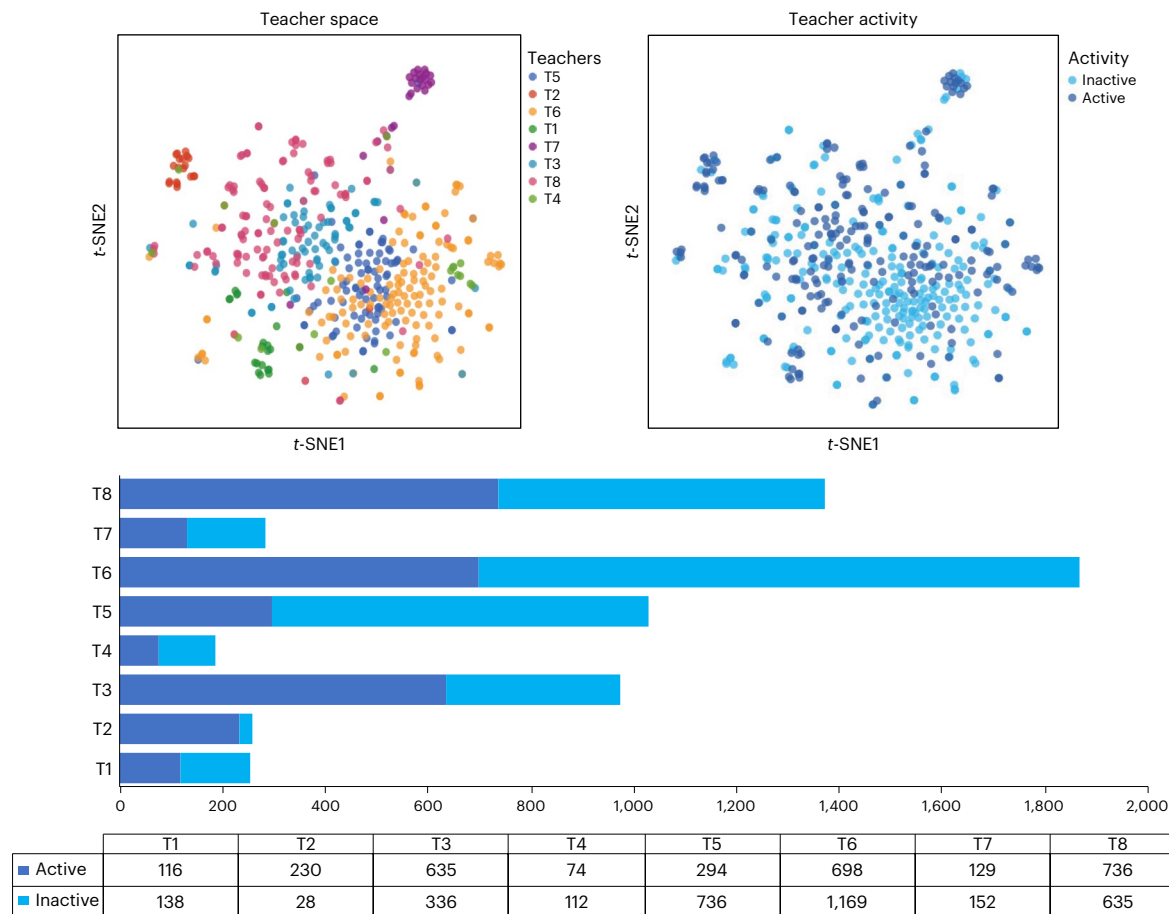


Fig. 6 | Public proof-of-concept datasets. Top left, the t -distributed stochastic neighbour embedding (t -SNE)⁶⁹ projection of the ChEMBL hERG data clusters used as virtual partner private data. Bottom left, the t -SNE projection of the

transfer, training and test chemical spaces. Top right, the distribution of the active and inactive compounds in the partner and teacher space. Bottom, the size and activity distribution of the individual partner data.

Both the public consortium simulation and the industrial FL experiment have demonstrated the successful extraction, transfer and federation of knowledge in the context of hERG activity data classification. We are confident that FLuID is an effective FL approach that can be used to facilitate knowledge sharing in the context of drug discovery.

Conclusions

We have introduced FLuID, a privacy-preserving approach to share knowledge currently locked in proprietary data silos and its application in the context of drug discovery. We have demonstrated FLuID's ability to extract, transfer and federate knowledge across multiple organizations without disclosing confidential information. The method was proven successful in validation experiments run on public data and in a real-world scenario using private pharmaceutical industry data. For the hERG binding classification task, we showed that the proposed FL approach allows us to substantially improve the performance and the AD of models when knowledge is shared between partners of a consortium. The gain in performance increases with the size of the consortium. The contribution of knowledge is distributed across all the participants, so all partners benefit from the collective knowledge and AD embedded in the original local private data.

Although the presented results are encouraging, it is important to note that in a real application context each partner operates in its own local domain that is likely to be more challenging than the public benchmark space used in our setup. Indeed, the chemical space of pharmaceutical companies is usually very focused and therefore does not reflect the public test space. This can explain the relatively

poor performance of the private teachers compared to the federated student or hybrid models that benefit from the more diverse federated data. In the real world, however, the model needs to perform well in global chemical space for new projects as well as in the more local space of each partner for existing projects. Only a subset of the knowledge contained in the federated data is transferable to the partner domain, the rest of the data might even constitute noise for a local model⁵¹. Hence this method requires the alignment of the federated knowledge with the target domain of a project. This well-known domain-shift challenge affects the performance of the hybrid model within each of the partners' spaces. Domain shift is a broad concern in ML not specific to FLuID. Thanks to the data-driven nature of FLuID and the large size of the federated dataset, it is possible to consider a wide spectrum of ML techniques to achieve domain adaptation and thus mitigate domain-shift issues. Such techniques include transfer learning, model fine tuning, adaptive data subsampling and so on. This exciting research theme is outside the scope of this work and will be reported in future communications.

The presented method is data driven as opposed to currently more popular model-centric FL. FLuID leverages knowledge distillation and introduces an implementation of FD in drug discovery. The use of surrogate data, KD and label aggregation ensures that labels remain anonymous and untraceable to the original training data, thereby exempting them from data protection and governance constraints. FLuID is algorithm agnostic and supports heterogeneous knowledge extraction across partners. The resulting knowledge captured in the form of annotated data can be subsequently used in many different ML

and AI workflows. The data-centric knowledge format is persistent and will benefit from future AI innovations. The underlying principle on the presented method is simple, intuitive and ensures privacy preservation while delivering promising knowledge transfer efficacy. The method enables access to a wealth of knowledge otherwise locked in private data. The resulting federated knowledge combined with private and public data opens an avenue to a whole new generation of statistical models with unprecedented performance and ADs in the context of biological activity prediction provided domain adaptation is achieved when necessary.

Methods

Knowledge extraction

The knowledge contained in the private data is extracted using an ML algorithm and captured in the form of a predictive model called the teacher model. Any appropriate algorithm combined with a relevant domain representation can be used to extract the knowledge. This operation requires direct access to the confidential training data and is therefore performed securely within the private space of the contributing organization.

Once the teacher model has been trained with private data, it is used to annotate a carefully selected collection of public domain chemical compounds. This annotation allows the knowledge embedded in the teacher model to be transferred and stored using the public structures. The resulting data do not contain confidential information and can leave the private space of the partner. The data can then be merged with the labelled data provided, in the same manner, by all the other organizations, forming a robust federated dataset.

The collection of public structures used to convey the knowledge is called the transfer data; it is a critical element of the method since it is responsible for capturing the knowledge embedded in the teacher model across a preferably wide AD. Hence, the transfer data require the following properties:

- **Diversity:** the transfer data must cover a large variety of domain examples (in our case, chemical structures) to effectively capture knowledge originating from different sources with probably different application domains.
- **Tractability:** although it would theoretically be desirable to use an extremely large size for the public transfer data, pragmatically it is necessary to limit the size with respect to storage, the prediction time required by the teacher model during the annotation phase, and the subsequent model training using the federated data.
- **Homogeneity:** the domain covered by the public transfer data should not privilege any arbitrary area. In other words, the transfer data should be distributed evenly across the target AD.

In our validation experiment, we selected 1 million diverse public structures randomly sampled from an initial pool of 100 million structures provided by the PubChem database⁵²; homogeneity was ensured by retaining only one representative structure per tile of chemical space. A tile defines a region of the chemical space delimited by a similarity radius around a centroid (reference) structure. In our case, the tiles have a radius of 0.5 Tanimoto⁵³ similarity when using the ECFP4 (Extended Connectivity Finger Print 4) (ref. 54) fingerprint representation of chemical structures. This sphere exclusion⁵⁵ protocol results in a collection of roughly 350,000 structures covering evenly the wide PubChem chemical space. The design of the transfer dataset is very important and can be adapted to different types of objective. For instance, it can be customized to federated knowledge within a specific target domain (drug space, cosmetic space, agrochemical space and so on).

Knowledge consolidation

The knowledge embedded in the transfer data collected across all contributing organizations is consolidated into a single federated dataset.

Within the private space of each organization, each teacher model contributed a hERG activity label for each of the 350,000 public compounds. These labels do not disclose the private data used to build the teacher model and can therefore be collected by a trusted third-party organization (honest broker) to perform the consolidation process. Each individual teacher prediction contains the following information:

- A label in the form of the likelihood for each category (ACTIVE or INACTIVE) in the form of a values between 0 and 1.0.
- A reliability level based on the average similarity between the public query compound and the k nearest neighbours ($k = 8$ by default) in the teacher's training set. For instance, using a Tanimoto⁵³ similarity based on ECFP representation^{54,56}. The reliability level expresses the relevance of the data supporting the prediction based on the structural proximity between the query instance and the supporting examples in the training data. The closer and therefore more relevant the training examples are to the query instance, the more reliable the prediction is under the chemical similarity principle^{57,58}.

The consolidation step offers an opportunity to design effective heuristics to merge the labels and maximize the synergistic effect of the ensemble of independent predictions. Optimizing the consolidation strategy is an interesting research topic, deserving its own attention and is outside the scope of this article. For the sake of this proof-of-concept demonstration, we choose a very simple and intuitive weighted average of the category likelihoods using the reliability as the weighting factor (equation (1)).

$$L_f = \frac{\sum_{i=1}^P r_i \times L_i}{\sum_{i=1}^P r_i} \quad (3)$$

where L_f is the resulting federated label, P is the number of partners in the consortium, L_i the label of the prediction provided by the i th teacher and r_i is the associated reliability.

With this simple approach, we preserve the information contained in the likelihood distribution of each label while moderating the impact of this label with the proximity of the public compound to the chemical space of the corresponding teacher. In other words, the closer a public compound is to the chemical space of a teacher, the more the label of this teacher will contribute to the federated label. The resulting federated label benefits from the knowledge of all contributing teachers that convey the knowledge embedded in the original private data. Label consolidation is a key component of DD-FL.

The federated dataset based on the consolidated labels becomes a new source of knowledge and is used to build a secondary model called the student model. As before, we can use any suitable modelling setup for this task, which means that FLuID supports heterogeneous FL at both the extraction and exploitation ends. For the sake of comparison in this work, we used the same algorithm, descriptors and model configuration to build all the models within each experiment. The predictive performance of the federated student model allows us to measure the depth of the shared knowledge, and its AD gives us a measure of the breadth of this knowledge.

Knowledge fusion

Although the federated student model is the central focus of this work, the most beneficial step for the federation partners is the integration of the federated knowledge into their internal model building process. Typically, this is achieved through the combination of federated data with private data (and potentially public data) to build powerful hybrid models. FLuID does not impose any modelling setup. Therefore, each partner may use bespoke approaches to optimally leverage the federated knowledge. For instance, the federated data can be merged with the private data to build a new model or used separately to pretrain a

model that can then be refined with private data using transfer learning or fine tuning techniques³⁹. This flexibility is an important feature of FLuID; the format of the resulting federated knowledge is compatible with experimental data and is easily integrated in various ML or AI workflows.

Federated data selection

Since the federated dataset is very large and diverse, we can take advantage of this size to select a subset for a given use case by applying filters to exploit only the most relevant federated data. For instance, one could prioritize confident data to create robust global models or focus on the target domain to produce more localized models. It is also possible to modulate the bias of the training set towards a given class or prioritize sensitivity over precision and vice versa. The ideal selection criteria for the subset of federated data depends on the use case and varies with the prediction task, the chosen learning algorithm, domain representation and, finally, the desired AD. Such flexibility is a powerful feature specific to the data-driven approach. It allows to optimize, a posteriori, on how the federated knowledge is consumed. In comparison, MD-FL knowledge is locked in a rigid model with limited adaptability.

For the sake of simplicity in our experiment, we have chosen to select a random subset of 10,000 datapoints composed of 5,000 compounds in each classification category (ACTIVE or INACTIVE), which leads to a perfectly balanced training set from which to build the student model.

Public proof of concept

We used public hERG binding data from ChEMBL⁶⁰ as our reference chemical space; after curation, it comprises 7,772 classification instances. The dataset was subsequently randomly split into the global training set D_{train} (80%, $n = 6,218$) and a validation set D_{val} (20%, $n = 1,554$); the latter was used for internal validation purpose reported in the notebook experiment.

Next, the global training set D_{train} was split into eight subsets covering different regions of the chemical space, using a k -means⁶¹ clustering algorithm ($k = 8$). This allows us to simulate a consortium of eight virtual partners with each cluster representing a virtual private data space used to train the teacher models (T1–T8). The cluster datasets have their own specific size, category bias and chemical space, mimicking a real-world situation. The 350,000 structures from the transfer data occupy a wide chemical space that covers well both the training and test domains (Fig. 6).

Individual teachers (T1–T8) were trained using the corresponding data. The resulting models were then tested against an external hERG benchmark dataset (D_{test}) published by Preissner et al.⁶² (duplicates between training and test set were removed from the test set). The test set consists of 4,383 compounds, 3,299 inactive (75%) and 1,084 actives (25%) and is therefore biased towards the inactive category.

Next, the 350,000 public structures were annotated by each teacher, leading to eight labels, L1–L8, per structure. Using the weighted average method described earlier (equation (1)), the eight labels, L1–L8 were consolidated into a single federated label L_{fed} for each structure resulting in the federated dataset D_{fed} .

Finally, 5,000 random ACTIVE compounds and 5,000 random INACTIVE compounds were selected from D_{fed} and used to train the student model S_{fed} .

Like the teachers, the resulting federated student model was evaluated against the external Preissner et al. benchmark data (D_{test}). The performance of the student was then compared to the performance of the individual teachers, T1–T8. The average teacher performance was finally compared with the federated student performance to validate our knowledge transfer hypothesis.

To build both the teacher and the student models, we used a random forest⁶³ learning algorithm and ECFP⁶⁶ as the representation for the chemical structure. This combination was chosen to match a widely

known methodology across the QSAR modelling community.

In the public validation experiment we recombined the federated data with the private data of each virtual partner and used this fused data to train hybrid models, H1–H8. The performance of the hybrid models was compared with the corresponding original teacher model performance to measure the benefits of sharing knowledge with the proposed FLuID approach. Alongside predictive performance, it is critical to consider the AD of a model. We therefore compared the AD of each teacher with its corresponding hybrid model to quantify the expected AD expansion. We used TARDIS⁶⁴ as the reference AD framework that defines compounds as out of AD if they contain a new feature never seen (absent) in the training data. During validation, AD coverage is expressed as the ratio of the predicted data falling inside the AD. Since the transfer data were designed to cover a large AD, we anticipated the student and hybrid models would exhibit an augmented AD compared to the individual teachers (training with focused cluster data).

Additionally, we demonstrate that knowledge has been successfully transferred within the newly applicable chemical space. For that purpose, we compared the performance of the teachers and their corresponding hybrid models for the subset of compounds that were outside the AD of the teacher model but brought inside the AD of the hybrid model. An increased performance in those compounds would show effective AD augmentation (coverage and predictivity).

Beyond comparing the predictive performance and AD characteristics of the teacher, student and hybrid models, we also investigated the effect of the size of the student model on its performance to gauge the density of knowledge in the federated data.

We considered several additional aspects of FLuID, which are important for the success of an FL consortium:

- Relative contribution of each partner to the federated knowledge
- Reward and/or impact of individual partners on the quality of the federated knowledge
- Impact of the number of partners in the consortium

Application in a real-world industrial context

We repeated the above proof-of-concept protocol in the context of the pharmaceutical industry by setting up a collaborative research consortium (Sanofi, Merck KGaA, Bayer AG, F. Hoffmann-La Roche AG, Novartis AG, UCB, Takeda, GSK and AstraZeneca), and ran the validation experiment across all eight partners in this consortium. On average, participants built their teacher using ~10,000 private instances with a range of contributions from 3,000 to 70,000 instances.

In the industrial setup, we built both the teacher and the student models using self-organizing hypothesis networks⁶⁵ as the learning algorithm. This algorithm offers good predictive performance while remaining transparent and interpretable. We used a pharmacophoric atom-pair-based fingerprint^{66,67} to represent the chemical structures. This combination of algorithm and descriptor is the result of investigations conducted in our research team for modelling receptor-based endpoints and displayed good performance when modelling the hERG endpoint using private pharmaceutical data⁶⁷. To measure the success of knowledge transfer, we used the same external hERG Preissner et al. benchmark D_{test} as for the simulation experiment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data for the public simulation experiment are held within the GitHub repository (https://github.com/LhasaLimited/FLuID_POC) in the data folder. A release of the repository is also available from Zenodo⁶⁸ (<https://zenodo.org/records/14531198>).

Code availability

All code necessary to run the public portion of the experiment can be found in the github repository (https://github.com/LhasaLimited/FLuID_POC) or it is also available from Zenodo⁶⁸ (<https://zenodo.org/records/14531198>). Simply open and run the Jupyter notebook and it will run the entire experiment using the included data. The code is licenced using the GPLv3 licence.

References

- Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Health* **1**, e271–e297 (2019).
- Zhou, W. et al. Ensembled deep learning model outperforms human experts in diagnosing biliary atresia from sonographic gallbladder images. *Nat. Commun.* **12**, 1259 (2021).
- Topaloglu, M. Y., Morrell, E. M., Rajendran, S. & Topaloglu, U. In the pursuit of privacy: the promises and predicaments of federated learning in healthcare. *Front. Artif. Intell.* **4**, 746497 (2021).
- Brauneck, A. et al. Federated machine learning in data-protection-compliant research. *Nat. Mach. Intell.* **5**, 2–4 (2023).
- Bak, M. et al. Federated learning is not a cure-all for data ethics. *Nat. Mach. Intell.* **6**, 370–372 (2024).
- Zhu, H., Xu, J., Liu, S. & Jin, Y. Federated learning on non-IID data: a survey. *Neurocomputing* **465**, 371–390 (2021).
- McMahan, B., Moore, E., Ramage, D., Hampson, S. & Arcas, B. A. y. Communication-efficient learning of deep networks from decentralized data. In *Proc. 20th International Conference on Artificial Intelligence and Statistics PMLR* **54**, 1273–1282 (2017).
- Zhou, J. et al. A survey on federated learning and its applications for accelerating industrial internet of things. Preprint at <https://doi.org/10.48550/arXiv.2104.10501> (2021).
- Li, L., Fan, Y., Tse, M. & Lin, K.-Y. A review of applications in federated learning. *Comput. Ind. Eng.* **149**, 106854 (2020).
- Li, T., Sahu, A. K., Talwalkar, A. & Smith, V. Federated learning: challenges, methods, and future directions. *IEEE Signal Process. Mag.* **37**, 50–60 (2020).
- Yin, X., Zhu, Y. & Hu, J. A comprehensive survey of privacy-preserving federated learning: a taxonomy, review, and future directions. *ACM Comput. Surv.* **54**, 131:1–131:36 (2021).
- Kairouz, P. et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning* **14**, 1–210, (2021).
- Liu, J. et al. From distributed machine learning to federated learning: a survey. *Knowl. Inf. Syst.* **64**, 885–917 (2022).
- Konečný, J., McMahan, H. B., Ramage, D. & Richtárik, P. Federated optimization: distributed machine learning for on-device intelligence. Preprint at <https://doi.org/10.48550/arXiv.1610.02527> (2016).
- Abadi, M. et al. Deep learning with differential privacy. In *Proc. 2016 ACM SIGSAC Conference on Computer and Communications Security* 308–318 (ACM, 2016).
- Dwork, C. Differential privacy: a survey of results. In *Proc. International Conference on Theory and Applications of Models of Computation* (eds Agrawal, M. et al.) 1–19 (Springer, 2008).
- Long, G., Tan, Y., Jiang, J. & Zhang, C. in *Federated Learning: Privacy and Incentive* (eds Yang, Q. et al.) 240–254 (Springer, 2020).
- Rieke, N. et al. The future of digital health with federated learning. *Npj Digit. Med.* **3**, 119 (2020).
- Choudhury, O. et al. Predicting adverse drug reactions on distributed health data using federated learning. *AMIA. Annu. Symp. Proc.* **2019**, 313–322 (2020).
- Nguyen, D. C. et al. Federated learning for smart healthcare: a survey. *ACM Computing Surveys (Csur)* **55**, 1–37 (2022).
- Xiong, Z. et al. Facing small and biased data dilemma in drug discovery with enhanced federated learning approaches. *Sci. China Life Sci.* **65**, 529–539 (2022).
- Manu, D. et al. FL-DISCO: federated generative adversarial network for graph-based molecule drug discovery: special session paper. In *Proc. 2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)* 1–7 (IEEE, 2021).
- Naz, S., Phan, K. T. & Chen, Y.-P. P. A comprehensive review of federated learning for COVID-19 detection. *Int. J. Intell. Syst.* **37**, 2371–2392 (2022).
- Goldsmith, M. R. et al. in *Crop Protection Products for Sustainable Agriculture* (eds Rauzan, B. M. & Lorschbach, B. A.) Vol. 1390, 181–200 (American Chemical Society, 2021).
- Heyndrickx, W. et al. MELLODDY: cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. *J. Chem. Inf. Model.* **64**, 2331–2344 (2024).
- Hanser, T. Federated learning for molecular discovery. *Curr. Opin. Struct. Biol.* **79**, 102545 (2023).
- Konečný, J. et al. Federated learning: strategies for improving communication efficiency. Preprint at <https://doi.org/10.48550/arXiv.1610.05492> (2017).
- Wu, C., Wu, F., Lyu, L., Huang, Y. & Xie, X. Communication-efficient federated learning via knowledge distillation. *Nat. Commun.* **13**, 2032 (2022).
- Zhu, X. *Semi-Supervised Learning Literature Survey* (Univ. Wisconsin, 2005); <https://minds.wisconsin.edu/handle/1793/60444>
- Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. Preprint at <http://arxiv.org/abs/1503.02531> (2015).
- Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I. & Talwar, K. Semi-supervised knowledge transfer for deep learning from private training data. Preprint at <http://arxiv.org/abs/1610.05755> (2016).
- Papernot, N. et al. Scalable private learning with PATE. Preprint at <http://arxiv.org/abs/1802.08908> (2018).
- Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
- Dietterich, T. G. Ensemble methods in machine learning. In *Proc. International Workshop on Multiple Classifier Systems* (eds Kittler, J. & Roli, F.) 1–15 (Springer, 2000).
- Li, L., Gou, J., Yu, B., Du, L. & Tao, Z. Y. D. Federated distillation: a survey. Preprint at <https://doi.org/10.48550/arXiv.2404.08564> (2024).
- Eldar, Y. C. et al. in *Machine Learning and Wireless Communications* (eds Goldsmith, A. et al.) 457–485 (Cambridge Univ. Press, 2022).
- Li, D. & Wang, J. FedMD: heterogenous federated learning via model distillation. Preprint at <https://doi.org/10.48550/arXiv.1910.03581> (2019).
- Itahara, S., Nishio, T., Koda, Y., Morikura, M. & Yamamoto, K. Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-IID private data. *IEEE Trans. Mob. Comput.* **22**, 191–205 (2023).
- Sattler, F., Marban, A., Rischke, R. & Samek, W. Communication-efficient federated distillation. Preprint at <http://arxiv.org/abs/2012.00632> (2020).
- Sui, D. et al. FedED: federated learning via ensemble distillation for medical relation extraction. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (eds Webber, B. et al.) 2118–2128 (Association for Computational Linguistics, 2020).

41. Han, S. et al. FedX: unsupervised federated learning with cross knowledge distillation. In *European Conference on Computer Vision*. (eds Avidan, S. et al.) 691–707 (Springer Nature Switzerland, 2022).
42. Jeong, E. et al. Communication-efficient on-device machine learning: federated distillation and augmentation under non-IID private data. Preprint at <http://arxiv.org/abs/1811.11479> (2023).
43. Choquette-Choo, C. A. et al. CaPC learning: confidential and private collaborative learning. Preprint at <https://doi.org/10.48550/arXiv.2102.05188> (2021).
44. PyGrid: a peer-to-peer platform for private data science and federated learning *OpenMined Blog* <https://blog.openmined.org/what-is-pygrid-demo/> (2020).
45. FLUID POC platform. *GitHub* https://github.com/LhasaLimited/FLUID_POC (2023).
46. Hancox, J. C., McPate, M. J., El Harchi, A. & Zhang, Y. H. The hERG potassium channel and hERG screening for drug-induced torsades de pointes. *Pharmacol. Ther.* **119**, 118–132 (2008).
47. Wolford, B. What is GDPR, the EU's new data protection law? *GDPR.eu* <https://gdpr.eu/what-is-gdpr/> (2018).
48. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. Membership inference attacks against machine learning model. In *IEEE Symposium on Security and Privacy (SP)* 3–18 (IEEE, 2017).
49. Raipuria, G., Bonthu, S. & Singhal, N. Noise robust training of segmentation model using knowledge distillation. In *Proc. Pattern Recognition. ICPR International Workshops and Challenges. ICPR 2021* (eds Del Bimbo, A. et al.) 97–104 (Springer, 2021).
50. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442–451 (1975).
51. Bassani, D., Brigo, A. & Andrews-Morger, A. Federated learning in computational toxicology: an industrial perspective on the Effiris Hackathon. *Chem. Res. Toxicol.* **36**, 1503–1517 (2023).
52. Kim, S. et al. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **49**, D1388–D1395 (2021).
53. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminformatics* **7**, 20 (2015).
54. Glen, R. et al. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs Investig. Drugs J.* **9**, 199–204 (2006).
55. Hudson, B. D., Hyde, R. M., Rahr, E., Wood, J. & Osman, J. Parameter based methods for compound selection from chemical databases. *Quant. Struct. Act. Relatsh.* **15**, 285–289 (1996).
56. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
57. Maggiora, G., Vogt, M., Stumpfe, D. & Bajorath, J. Molecular similarity in medicinal chemistry. *J. Med. Chem.* **57**, 3186–3204 (2014).
58. Maggiora, G. M. *Concepts and Applications of Molecular Similarity* (eds Johnson, M. A. & Maggiora, G. M.) (John Wiley & Sons, 1990).
59. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
60. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).
61. MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Symp. Math. Stat. Probab.* (eds Marie Le Cam, L. & Neyman, J.) Vol. 1, 281–298 (1967).
62. Siramshetty, V. B., Chen, Q., Devarakonda, P. & Preissner, R. The catch-22 of predicting hERG blockade using publicly accessible bioactivity data. *J. Chem. Inf. Model.* **58**, 1224–1233 (2018).
63. Ho, T. K. Random decision forests. In *Proc. 3rd International Conference on Document Analysis and Recognition* Vol. 1, 278–282 (1995).
64. Hanser, T., Barber, C., Marchaland, J. F. & Werner, S. Applicability domain: towards a more formal definition. *SAR QSAR Environ. Res.* <https://doi.org/10.1080/1062936X.2016.1250229> (2016).
65. Hanser, T. et al. Self organising hypothesis networks: a new approach for representing and structuring SAR knowledge. *J. Cheminformatics* **6**, 21 (2014).
66. Carhart, R., Smith, D. H. & Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* <https://doi.org/10.1021/ci00046a002> (1985).
67. Hanser, T., Steinmetz, F. P., Plante, J., Rippmann, F. & Krier, M. Avoiding hERG-liability in drug design via synergetic combinations of different (Q)SAR methodologies and data sources: a case study in an industrial setting. *J. Cheminformatics* **11**, 9 (2019).
68. Hanser, T., Werner, S. & Plante, J. FLUID POC a simulation platform for federated distillation. *Zenodo* <https://doi.org/10.5281/zenodo.14531198> (2024).
69. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

Acknowledgements

The Lhasa Limited authors acknowledge the invaluable collaborative effort, scientific discussions and knowledge sharing made possible by all the industrial partners in this project.

Author contributions

E.A., A.A., L.T.A., R.J.B., A.B., A.D., S.G., N.G., D.K., L.K., W.M., F.R., Y.S., F.S., A.W., J.W. and T.Y.: data access and preparation, domain expertise and continuous input and knowledge sharing. T.H., J.-F.M., J.P., R.v.D. and S.W.: methodology design and implementation, experiment orchestration, virtual simulation platform development, data analytics and paper preparation. C.B. and L.J.: managing the partner–Lhasa relationships.

Competing interests

The authors declare no competing interests, except for R.J.B. who is an employee and shareholder of Sanofi, a pharmaceutical R&D company that may benefit from the outcome of this research.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42256-025-00991-2>.

Correspondence and requests for materials should be addressed to Thierry Hanser.

Peer review information *Nature Machine Intelligence* thanks Alissa Brauneck, Gabriele Buchholtz, Stuart McLennan and Umit Topaloglu for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2025

¹Lhasa Limited, Leeds, UK. ²Predictive Compound ADME and Safety, Drug Safety and Metabolism, AstraZeneca IMED Biotech Unit, Mölndal, Sweden. ³Mölnlycke Health Care, Gothenburg, Sweden. ⁴Preclinical Safety, Sanofi, Frankfurt, Germany. ⁵Computational Toxicologist, Genentech, Inc., South San Francisco, CA, USA. ⁶Preclinical Safety, Sanofi, Cambridge, MA, USA. ⁷Pharma Research and Early Development, Roche, Basel, Switzerland. ⁸Safety and Secondary Pharmacology, UCB Biopharma SRL, Braine L'Alleud, Belgium. ⁹Novartis Institutes for Biological Research (NIBR), Novartis Pharma AG, Basel, Switzerland. ¹⁰Imaging & Data Analytics, Clinical Pharmacology & Safety Sciences, AstraZeneca, Waltham, MA, USA. ¹¹Toxicology, Recursion Pharmaceuticals, Salt Lake City, UT, USA. ¹²Global Computational Chemistry and Biology, Merck Healthcare KGaA, Darmstadt, Germany. ¹³Pharmaceuticals, R&D, Computational Molecular Design, Bayer AG, Berlin, Germany. ¹⁴AI & Modeling, DSM-Firmenich AG, Geneva, Switzerland. ¹⁵Computational Toxicology, GlaxoSmithKline (GSK), Stevenage, UK. ¹⁶Computational Chemistry, Selvita S.A, Cracow, Poland. ¹⁷Chemical Toxicology, Takeda Pharmaceutical Company, Ltd, Fujisawa, Japan. ✉e-mail: thierry.hanser@lhasalimited.org

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
<input type="checkbox"/>	<input checked="" type="checkbox"/> The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement
<input type="checkbox"/>	<input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
<input checked="" type="checkbox"/>	<input type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
<input checked="" type="checkbox"/>	<input type="checkbox"/> A description of all covariates tested
<input type="checkbox"/>	<input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
<input type="checkbox"/>	<input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
<input checked="" type="checkbox"/>	<input type="checkbox"/> For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
<input checked="" type="checkbox"/>	<input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
<input checked="" type="checkbox"/>	<input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
<input checked="" type="checkbox"/>	<input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	A complete description of the data collection and preparation can be found in the open access Jupyter Notebook at: https://github.com/LhasaLimited/FLuID_POC (2023). In the manuscript this is referred to as '. For transparency purpose, the method evaluation is based on public data and available as open-source software49.' We have run the notebook using Jupyter version 7.0.8 but any current version should work.
Data analysis	A complete description of the data analysis and interpretation can be found in the open access Jupyter Notebook at: https://github.com/LhasaLimited/FLuID_POC (2023). In the manuscript this is referred to as '. For transparency purpose, the method evaluation is based on public data and available as open-source software49.'

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

A complete description of the data accessibility and code to read access the data can be found in the open access Jupyter Notebook at: https://github.com/LhasaLimited/FLuID_POC (2023). In the manuscript this is referred to as '. For transparency purpose, the method evaluation is based on public data and available as open-source software⁴⁹.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender

N/A

Population characteristics

N/A

Recruitment

N/A

Ethics oversight

N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Life science is being used as it is the most applicable choice from the above three choices, but most of these concerns aren't applicable to a computational study. In terms of the sample size, we collected all of the data that we deemed complete from ChEMBL. When referring to N in the manuscript it refers to running exact replicates to account for the inherent variability with machine learning not with regards to selecting a sample size. The manuscript itself discusses the splitting of the training data into groups to simulate different companies, but it is one of the points of the manuscript and discussed there

Data exclusions

The only data excluded was during the building of the initial dataset from ChEMBL. When gathering data containing hERG activity inconclusive data points were removed along with any data that could not be resolved into an active or inactive call.

Replication

To reproduce the findings simply clone the github repository and run the Jupyter notebook. The exact numbers will not be identical, but they will be within acceptable variation and the minor randomness inherent with machine learning. Nothing was done in actual living creatures that would require replication or a sample size.

Randomization

No data was randomized. We did ensure that any compounds from the test set (obtained from a reference paper) that were present in the training set were removed from the test set.
Otherwise there was no splitting other than what was described in the paper, which is one of the points of the paper and discussed there

Blinding

Blinding isn't appropriate as this is a computational experiment. The computer will not initiate any bias requiring a double blind study to eliminate. It instead will apply statistics mathematically to the data and thus no blinding is required.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging